Statistics as Inductive Inference

Jan-Willem Romeijn
Faculty of Philosophy
University of Groningen
j.w.romeijn@rug.nl

Abstract

This chapter[1] concerns the relation between statistics and inductive logic. I start by describing induction in formal terms, and I introduce a general notion of probabilistic inductive inference. This provides a setting in which statistical procedures and inductive logics can be captured. Specifically, I discuss three statistical procedures (hypotheses testing, parameter estimation, and Bayesian statistics) and I show to what extend they can be captured by certain inductive logics. I end with some suggestions on how inductive logic can be developed so that its ties with statistics are strengthened.

1 Statistical procedures as inductive logics

An inductive logic is a system of inference that describes the relation between propositions on data, and propositions that extend beyond the data, such as predictions over future data, and general conclusions on all possible data. Statistics, on the other hand, is a mathematical discipline that describes procedures for deriving results about a population from sample data. These results include predictions on future samples, decisions on rejecting or accepting a hypothesis about the population, the determination of probability assignments over such hypotheses, the selection of a statistical model for studying the population, and so on. Both inductive logic and statistics are calculi for getting from the given data to propositions or results that transcend the data.

[1] This chapter was written in parallel with my chapter for the Handbook for the History of Logic: Inductive Logic (Volume 10), edited by Hartmann et al. (2009). The two chapters show a considerable overlap. The present chapter aims at a reconstruction of statistical procedures in terms of inductive logics. The other chapter approaches the same material from the other end, and considers how inductive logic can be developed to encompass the statistical procedures.

This suggests that there is a strong parallel between statistics and inductive logic. In fact, it does not take much imagination to view statistical procedures as inferences: the input components, primarily the data, are the premises, and the result of the procedure is the conclusion. In this rough and ready way, statistical procedures can be understood as defining particular inductive logics. However, the two disciplines have evolved more or less separately. In part this is because there are objections to viewing classical statistics as inferential, although this is not true for all statistical procedures. For another part, it may be because inductive logic has been dominated by the Carnapian programme. Perhaps statisticians have not recognised inductive logic as a discipline that is much like their own.

However this may be, I think it is time for a rapprochement. There are, to my mind, good reasons for investigating the parallel between inductive logic and statistics along the lines suggested above. First, framing the statistical procedures as inferences in a logic may clarify the presuppositions of these procedures. Second, by relating statistics to inductive logic, techniques and insights from inductive logic may be used to enrich statistics. And finally, showing the parallels between inductive logic and statistics may show the relevance, also to inductive logicians themselves, of their discipline to the sciences, and thereby direct further research in this field.

With this aim in mind, I consider a number of statistical procedures in this chapter, and I investigate whether they can be seen as part of an inductive logic, or otherwise whether they can, at least partly, be translated into such a logic. I start by describing induction in formal terms, and I introduce a general notion of probabilistic inductive inference. This provides a setting in which both statistical procedures and inductive logics may be captured. I subsequently discuss a number of statistical procedures, and show how they can, and cannot, be captured by certain inductive logics.

The first statistical procedure is Neyman-Pearson hypotheses testing. This procedure was introduced as explicitly non-inferential, and so it should strictly speaking not be captured by an inductive logic. On the other hand, power and significance are often interpreted inferentially. At the end of the chapter I devise an inductive logic that may be used to warrant such an interpretation. The second statistical procedure is parameter estimation. I briefly discuss Fisher's theory of maximum likelihood estimators, and I show that there is a certain relation with the inductive logic developed by Carnap. A third statistical procedure is Bayesian statistics. I show that it can be captured in a probabilistic inductive logic that relates to Carnapian inductive logic via the representation theorem of de Finetti. This leads to a discussion of Bayesian statistics in relation to Bayesian inductive logic.

Given the nature of the chapter, the discussion of statistical procedures is relatively short. Many procedures cannot be dealt with. Similarly, I cannot discuss in detail the many inductive logics devised within Carnapian inductive logic. For the former, the reader may consult other chapters in this volume, in particular the chapter by Festa. For the latter, I refer to Hartmann et al. (2009), specifically the discussions of inductive logic contained therein.

2 Observational data

As indicated, inductive inference starts from propositions on data, and ends in propositions that extend beyond the data. An example of an inductive inference is that, from the proposition that up until now all observed pears were green, we conclude that the next few pears will be green as well. Another example is that from the green pears we have seen we conclude that all pears are green, period. The key characteristic is that the conclusion says more than what is classically entailed by the premises.

Let me straighten these inferences out a bit. First, I restrict attention to propositions on empirical facts, thus leaving aside such propositions as that pears are healthy, or that God made them. Second, I focus on the results of observations of particular kinds of empirical fact. For example, the empirical fact at issue is the colour of pears, and the results of the observations are therefore colours of individual pears. There can in principle be an infinity of such observation results, but what I call data is always a finite sequence of them. Third, the result of an observation is always one from a designated partition of properties, usually finite but always countable. In the pear case, it may be {red, green, yellow}. I leave aside observations that cannot be classified in terms of a mutually exclusive set of properties.

I now make these ideas on what counts as data a bit more formal. The concept I want to get across is that of a sample space, in which single observations and sequences of observations can be represented as sets, called events. After introducing the observations in terms of a language, I define sample space. All the probabilities in this chapter will be defined over sample space because, probability is axiomatized as a measure function over sets. However, the expressions may be taken as sentences from a logical language just as well.

We denote the observation of individual i by Q_i. This is a propositional variable, and we denote assignments or valuations of this variable by q_i^k, which represents the sentence that the result of observing individual i is the property k. A sequence of such results of length t, starting at 1, is denoted with the propositional variable S_t, and its assignment with $s^{k_1 \dots k_t}$, often abbreviated as s_t. In order to simplify notation, I denote properties with natural numbers, so $k \in K = \{0, 1, \dots, n-1\}$. For example, if the observations are the aforementioned colours of pears, then $n = 3$. I write red as 0, green as 1, and yellow as 2, so s^{012} says that the first three pairs were red, green, and yellow respectively. Note further that there are logical relations among the sentences, like $s^{012} \rightarrow q_2^1$. Together, the expressions s_t and q_i^k form the observation language.

Now we develop a set-theoretical representation of the observations, a so-called sample space, otherwise known as an observation algebra. To this aim, consider the set of all infinitely long sequences K^ω, that is, all sequences like 012002010211112..., each encoding the observations of infinitely many pears. Denote such sequences with u, and write $u(i)$ for the i-th element in the sequence u. Every sentence q_i^k can then be associated with a particular set of such sequences, namely the set of u whose i-th element is k:

$$q_i^k = \{u \in K^\omega : \ u(i) = k\}.$$

Clearly, we can build up all finite sequences of results $s^{k_1 \dots k_t}$ as intersections of such sets:

$$s^{k_1 \dots k_t} = \bigcap_{i=1}^{t} q_i^{k_i}.$$

Note that entailments in the language now come out as set inclusions: we have $s^{012} \subset q_2^1$. Instead of a language with sentences q_i^k and logical relations among such sentences, I will in the following employ the algebra $\mathcal{Q}$, built up by the sets q_i^k and their conjunctions and intersections.

I want to emphasise that the notion of a sample space introduced here is really quite general. It excludes a continuum of individuals and a continuum of properties, but apart from that, any data recording that involves individuals and that ranges over a set of properties can serve as input. For example, instead of pears having colours we may think of subjects having test scores. Or of companies having certain stock prices. The sample space used in this chapter follows the basic structure of most applications in statistics, and of almost all applications in inductive logic.

3 Inductive inference

Now that I have made the notion of data more precise, let me turn to inductive inference. Consider the case in which I have observed three green pears: s^{111}. What can I conclude about the next pear? Or about pears in general? From the structure of the data itself, it seems that we can conclude depressingly little. We might say that the next pear is green, q_4^1. But as it stands, each of the sets $s^{111k} = s^{111} \cap q_4^k$, for $k = 0, 1, 2$, is a member of the sample space, or in terms of the logical language, we cannot derive any sentence q_4^k from the sentence s^{111}. The event of observing three green pears is consistent with any colour for the next pear. Purely on the basis of the classical relations among observations, as captured by the language and the sample space, we cannot draw any inductive conclusion.

Perhaps we can say that given three green pears, the next pear being green is more probable? This is where we enter the domain of probabilistic inductive logic. We can describe the complete population of pears by a probability function over the observational facts,

$$P : \mathcal{Q} \mapsto [0, 1].$$

Every possible pear q_{t+1}^k, and also every sequence of such pears $s^{k_1 \dots k_t}$, receives a distinct probability. The probability of the next pear being of a certain colour, conditional on a given sequence, is expressed as $P(q_{t+1}^k | s^{k_1 \dots k_t})$. Similarly, we may wonder about the probability that all pears are green, which is again determined by the probability assignment, in this case $P(\{\forall i : q_i^1\})$. All such probabilistic inductive inferences are determined by the full probability function P.

The central question of any probabilistic inductive inference or procedure is therefore how to determine the function P, relative to the data that we already have. What must the probability of the next observation be, given a sequence of observations gone before? And what is the right, or otherwise the preferable, distribution over all observations given the sequence? Both statistics and inductive logic aim to provide an answer to these questions, but they do so in different ways.

In order to facilitate the view that the statistical procedures are logical inferences, it will be convenient to keep in mind a particular understanding of probability assignments P over the sample space, or observation algebra, $\mathcal{Q}$. Recall that in classical two-valued logic, a model of the premises is a complete truth valuation over the language, subject to the rules of logic. Because of the correspondence between language and algebra, the model is also a complete function over the algebra, taking the values $\{0, 1\}$. Accordingly, the premises of some deductive logical argument are represented as a set of models over the algebra. By analogy, we may consider a probability function over an observation algebra as a model too. Just like the truth value assignment, the probability function is a function over an algebra, only it takes values in the interval $[0, 1]$, and it is subject to the axioms of probability.

Probabilistic inductive logics use probability models for the purpose of inductive inference. In particular, the premises of a probabilistic inductive argument can be represented as a set, possibly a singleton, of probability assignments. But there are widely different ways of understanding the inferential step, i.e., the step running from the premises to the conclusion. The most straightforward of these, and the one that is closest to classical statistical practice, is to associate a probability function P, or otherwise a set of such functions, with each sample s_t. The inferential step then runs from the data s_t and a large set of probability functions P, possibly all conceivable functions, towards a more restricted set, or even towards a single P. The resulting inductive logic is called *ampliative*, because the restriction on the set of probability functions that is effected by the data, i.e. the conclusion, is often stronger than what follows from the data and the initial set of probability functions, i.e. the premises, by deduction.

We can also make the inferential step precise by analogy to a more classical, *non-ampliative* notion of entailment. As will become apparent, this kind of inferential step is more naturally associated with what is traditionally called inductive logic. It is also associated with a basic kind of probabilistic logic, as elaborated in Hailperin (1996) and more recently in Haenni et al. (2009), especially section 2. Finally, this kind of inference is strongly related to Bayesian logic, as advocated by Howson (2003). It is the kind of inductive logic favored in this chapter.

An argument is said to be classically valid if and only if the set of models satisfying the premises is contained in the set of models satisfying the conclusion. The same idea of classical entailment may now be applied to the probabilistic models over sample space. In that case, the inferential step is from one set of probability assignments, characterised by a number of restrictions associated with premises, towards another set of probability assignments, characterised by a different restriction that is associated with a conclusion. The inductive inference is called valid if the former is contained in the latter, i.e., if every model satisfying the premises is also a model satisfying the conclusions. In such a valid inferential step, the conclusion does not amplify the premises.

As an example, say that we fix $P(q_1^0) = \frac{1}{2}$ and $P(q_1^1) = \frac{1}{3}$. Both these probability assignments can be taken as premises in a logical argument, and the models of these premises are simply all probability functions P over $\mathcal{Q}$ for which these two valuations hold. By the axioms of probability, we can derive that any such function P will also satisfy $P(q_1^2) = \frac{1}{6}$, and hence also that

$P(q_1^2) < \frac{1}{4}$. On its own, the latter expression amounts to a set of probability functions over the sample space $\mathcal{Q}$ in which the probability functions that satisfy both premises are included. In other words, the latter assignment is classically entailed by the two premises. Along exactly the same lines, we may derive a probability assignment for a statistical hypothesis h conditional on the data s_t, written as $P(h|s_t)$, from the input probabilities $P(h)$, $P(s_t)$, and $P(s_t|h)$, using the theorem of Bayes.

The classical, non-ampliative understanding of entailment may thus be used to reason inductively, towards predictions and statistical hypotheses that themselves determine a probability assignment over data. In the following I will focus primarily on such non-ampliative inductive logical inferences to investigate statistical procedures. I first discuss Neyman-Pearson hypothesis testing and Fisher's maximum likelihood estimation in their own terms, showing that they are best understood as ampliative inductive inferences. Then I discuss Carnapian inductive logic and show that it can be viewed as a non-ampliative version of parameter estimation. This leads to a discussion of Bayesian statistical inference, which is subsequently related to a generalisation of Carnapian inductive logic, Bayesian inductive logic. The chapter ends with an application of this logic to Neyman-Pearson hypothesis testing.

As indicated, Carnapian inductive logic is most easily related to non-ampliative logic. So, viewing statistical procedures in this perspective makes the latter more amenable to inductive logical analysis. But I do not want to claim that I thereby lay bare the real nature of the statistical procedures. Rather, I hope to show that the investigation of statistics along these specific logical lines clarifies and enriches statistical procedures. Furthermore, as indicated, I hope to stimulate research in inductive logic that is directed at problems in statistics.

4 Neyman-Pearson testing

The first statistical application concerns the choice between two statistical hypotheses, that is, two fully specified probability functions over sample space. In the above vocabulary, it concerns the choice between two probabilistic models, but we must be careful with our words here, because in statistics, models often refer to sets of statistical hypotheses. In the following, I will therefore refer to complete probability functions over the algebra as *hypotheses.*

Let $\mathcal{H} = \{h_0, h_1\}$ be the set of hypotheses, and let $\mathcal{Q}$ be the sample space, that is, the observation algebra introduced earlier on. We can compare the hypotheses h_0 and h_1 by means of a Neyman-Pearson test function. See Barnett (1999) and Neyman and Pearson (1967) for the details.

Definition 4.1 (Neyman-Pearson Hypothesis Test) *Let F be a function over the sample space $\mathcal{Q}$,*

$$F(s_t) = \begin{cases} 1 & \textit{if } \frac{P_{h_1}(s_t)}{P_{h_0}(s_t)} > r, \\ 0 & \textit{otherwise,} \end{cases} \tag{1}$$

where P_{h_j} is the probability over the sample space determined by the statistical hypothesis h_j. If $F = 1$ we decide to reject the null hypothesis h_0, else we accept h_0 for the time being.

Note that, in this simplified setting, the test function is defined for each set of sequences s_t separately. For each sample plan, and associated sample size t, we must define a separate test function.

The decision to accept or reject a hypothesis is associated with the so-called significance and power of the test:

$$\text{Significance}_F = \alpha = \int_{\mathcal{Q}} F(s_t) P_{h_0}(s_t) ds_t,$$

$$\text{Power}_F = 1 - \beta = \int_{\mathcal{Q}} F(s_t) P_{h_1}(s_t) ds_t.$$

The significance is the probability, according to the hypothesis h_0, of obtaining data that leads us to reject the hypothesis h_0, or in short, the type-I error of falsely rejecting the null hypothesis, denoted α. Similarly, the power is the probability, according to h_1, of obtaining data that leads us to reject the hypothesis h_0, or in short, the probability under h_1 of correctly rejecting the null hypothesis, so that $\beta = 1 - \text{Power}$ is the type-II error of falsely accepting the null hypothesis. An optimal test is one that minimizes the significance level, and maximizes the power. Neyman and Pearson prove that the decision has optimal significance and power for, and only for, likelihood-ratio test functions F. That is, an optimal test depends only on a threshold for the ratio $\frac{P_{h_1}(s_t)}{P_{h_0}(s_t)}$.

Let me illustrate the idea of Neyman-Pearson tests. Say that we have a pear whose colour is described by q^k, and we want to know from what farm it originates, from farmer Maria (h_0) or Lisa (h_1). We know that the colour composition of the pears from the two farms are different:

Hypothesis \ Data	q^0	q^1	q^2
h_0	0.00	0.05	0.95
h_1	0.40	0.30	0.30

If we want to decide between the two hypotheses, we need to fix a test function. Say that we choose

$$F(q^k) = \begin{cases} 0 & \text{if } k = 2, \\ 1 & \text{else.} \end{cases}$$

In the definition above, which uses a threshold for the likelihood ratio, this comes down to choosing a value for r somewhere between $\frac{6}{19}$ and 14, for example $r = 1$. The significance level is $P_{h_0}(q^0 \cup q^1) = 0.05$, and the power is $P_{h_1}(q^0 \cup q^1) = 0.70$. Now say that the pear we have is green, so $F = 1$ and we reject the null hypothesis, concluding that Maria did not grow the pear with the aforementioned power and significance.

Note that from the perspective of ampliative inductive logic, it is not too far-fetched to read an inferential step into the Neyman-Pearson procedure. The test function F brings us from a sample s_t and two probability functions, P_{h_j} for $j = 0, 1$, to a single probability function P_{h_1}, or P_{h_0}, over the sample space $\mathcal{Q}$. So we might say that the test function is the procedural analogue of an inductive inferential step, as discussed in Section 3. This step is ampliative because both probability functions P_{h_j} are consistent with the data. Ruling out one of them cannot be done deductively.[2]

Neyman-Pearson hypothesis testing is sometimes criticised because its results generally depend on the probability function over the entire sample space, and not just on the probability of those elements in sample space corresponding to the actual events, the observed sample for short. That is, the decision to accept or reject the null hypothesis against some alternative hypothesis depends not just on the probability of what has actually been observed, but also on the probability assignment over everything that could have been observed. A well-known illustration of this problem concerns so-called optional stopping. But here I want to illustrate the same point with

[2] There are attempts to make these ampliative inferences more precise, by means of a form of default reasoning, or a reasoning that employs a preferential ordering over probability models. Specifically, so-called evidential probability, proposed by Kyburg (1974) and more recently discussed by Wheeler (2006), is concerned with inferences that combine statistical hypotheses, which are each accepted with certain significance levels. However, in this chapter I will not investigate these logics. They are not concerned with inferences from the data to predictions or to hypotheses, but rather with inferences from hypotheses to other hypotheses, and from hypotheses to predictions.

an example that can be traced back to Jeffreys (1931) p. 357, and of which a variant is discussed in Hacking (1965).[3]

Instead of the hypotheses h_0 and h_1 above, say that we compare the hypotheses $h_0^\star$ and h_1.

Hypothesis \ Data	q^0	q^1	q^2
$h_0^\star$	0.05	0.05	0.90
h_1	0.40	0.30	0.30

We determine the test function $F(q^k) = 1$ iff $k = 0$, by requiring the same significance level, $P_{h_0^\star}(q^0) = 0.05$, resulting in the power $P_{h_1}(q^0) = 0.40$. Now imagine that we observe q^1 again. Then we accept $h_0^\star$. But this is a bit odd, because the hypotheses h_0 and $h_0^\star$ have the same probability for q^1! So how can the two test procedures react differently to this observation? It seems that, in contrast to h_0, the hypothesis $h_0^\star$ escapes rejection because it allocates some probability to s^0, an event that does not occur. This causes a shift in the area within sample space on which the hypothesis $h_0^\star$ is rejected. This phenomenon gave rise to the famed complaint of Jeffreys that "a hypothesis that may be true may be rejected because it has not predicted observable results that have not occurred".

This illustrates how the results of a Neyman-Pearson procedure depends on the whole probability function that a hypothesis defines over the sample space, and not just on the probability defined for the actual observation. From the perspective of an inductive logician, it may therefore seem "a remarkable procedure", to cite Jeffreys again. But it must be emphasised that Neyman-Pearson statistics was never intended as an inference in disguise. It is a procedure that allows us to decide between two hypotheses on the basis of data, generating error rates associated with that decision. Neyman and Pearson themselves were very explicit that the procedure must not be interpreted inferentially. Rather than inquiring into the truth and falsity of a hypothesis, they were interested in the probability of mistakenly deciding to reject or accept a hypothesis. The significance and power concern the probability over data given a hypothesis, not the probability of hypotheses given the data.[4]

[3] I would like to thank Jos Uffink for bringing this example to my attention. As far as I can see, the exact formulation of the example is his.

[4] This is not to say that Neyman-Pearson statistics cannot be viewed from an inferential angle. See Section 9 for an inferential account.

5 Fisher's parameter estimation

Let me turn to another important classical statistical procedure, so-called parameter estimation. I focus in particular on an estimation procedure first devised by Fisher, estimation by maximum likelihood.

The maximum likelihood estimator determines the best among a much larger, possibly infinite, set of hypotheses. Again it depends entirely on the probability that the hypotheses assign to points in the sample space. See Barnett (1999) and Fisher (1956) for more detail.

Definition 5.1 (Maximum Likelihood Estimation) *Let $\mathcal{H} = \{h_\theta : \theta \in \Theta\}$ be a set of hypotheses, labeled by the parameter θ, and let $\mathcal{Q}$ be the sample space. Then the maximum likelihood estimator of θ,*

$$\hat{\theta}(s_t) = \{\theta : \forall h_{\theta'} \big(P_{h_{\theta'}}(s_t) \leq P_{h_\theta}(s_t)\big)\}, \tag{2}$$

$\hat{\theta}$ for short, is a function over the elements s_t in the sample space.

So the estimator is a set, typically a singleton, of those values of θ for which the likelihood of h_θ on the data s_t is maximal. The associated best hypothesis we denote with $h_{\hat{\theta}}$. Note that this estimator is a function over the sample space, associating each s_t with a hypothesis, or a set of them.

Often the estimation is coupled to a so-called confidence interval. Restricting the parameter space to $\Theta = [0, 1]$ for convenience, and assuming that the true value is θ, we can define a region in sample space within which the estimator function is not too far off the mark. Specifically, we might set the region in such a way that it covers $1 - \epsilon$ of the probability P_{h_θ}:

$$\int_{\theta-\Delta}^{\theta+\Delta} P_{h_\theta}(\hat{\theta}) d\hat{\theta} = 1 - \epsilon.$$

We can provide an unproblematic frequentist interpretation of the interval $\hat{\theta} \in [\theta - \Delta, \theta + \Delta]$: in a series of estimations, the fraction of times in which the estimator $\hat{\theta}$ is further off the mark than Δ will tend to ϵ. The smaller the region, the more reliable the estimate. Note, however, that this interval is defined in terms of the unknown true value θ.

Some applications allow for the derivation of a region of parameter values within which the true value θ can be expected to lie.[5] The general idea is

[5] The determination of such regions is similar in nature to the determination of so-called fiducial probability. Fisher (1930, 1935, 1956) developed the notion of fiducial probability as a way of capturing parameter estimation in terms of a non-ampliative entailment relation, basically deriving a probability assignment over hypotheses without assuming a prior

to define a set of parameter values R within which the data are not too unlikely, $R(s_t) = \{\theta : P_{h_\theta}(s_t) > \epsilon\}$ for some small value $\epsilon > 0$. Now in terms of the integral above, we can swap the roles of θ and $\hat{\theta}$ and define the so-called central confidence interval:

$$\mathrm{Conf}_{1-\epsilon}(\hat{\theta}) = \left\{ \theta : |\theta - \hat{\theta}| < \Delta \text{ , and } \int_{\hat{\theta}-\Delta}^{\hat{\theta}+\Delta} P_{h_\theta}(\hat{\theta}) d\theta = 1 - \epsilon \right\}.$$

Via the function $\hat{\theta}(s_t)$, every element of the sample space s_t is assigned a region $\mathrm{Conf}_{1-\epsilon}$ of parameter values, interpreted as the region within which we may expect to find the true value θ. Note, however, that swapping the roles of θ and $\hat{\theta}$ in the integral is not unproblematic. We can only interpret the integral as a probability if $P_{h_\theta}(\hat{\theta} + \delta) = P_{h_{\theta-\delta}}(\hat{\theta})$ for all values of δ, or in other words, if for fixed $\hat{\theta}$ the function $P_{h_\theta}(\hat{\theta})$ is indeed a probability density over θ. In other cases, the interval cannot be taken as expressing the expected accuracy of the estimate, or at least not without further critical reflection.

Let me illustrate parameter estimation in a simple example on pears again. Say that we are interested in the colour composition of pears from Emma's farm, and that her pears are red, q_i^0, or green, q_i^1. Any ratio between these two kinds of pears is possible, so we have a set of hypotheses h_θ, called multinomial hypotheses, for which

$$P_{h_\theta}(q_t^1|s_{t-1}) = \theta, \qquad P_{h_\theta}(q_t^0|s_{t-1}) = 1 - \theta \tag{3}$$

with $\theta \in [0, 1]$. The hypothesis h_θ fixes the portion of green pears at θ, and therefore, independently of what pears we saw before, on the assumption of the hypothesis h_θ the probability that a randomly drawn pear from Emma's farm is green is θ. The type of distribution over $\mathcal{Q}$ that is induced by these hypotheses is called a Bernoulli distribution, or a multinomial distribution.

The idea of Fisher's maximum likelihood estimation is that we choose the value of θ for which the probability that the hypotheses h_θ gives to the data $s^{k_1 \dots k_t}$ is maximal. Say that we have observed a sequence of pears s^{000101}. The probability of these data given the hypothesis h_θ is

$$P_{h_\theta}(s^{000101}) \;=\; \prod_{i=1}^{t} P_{h_\theta}(q_i^{k_i}|s_{i-1}) = \theta^2(1-\theta)^4. \tag{4}$$

probability over statistical hypotheses at the outset. The fiducial argument is controversial, however, and its applicability is limited to particular statistical problems. Seidenfeld (1979) provides a detailed discussion of the restricted applicability of the argument in cases with multiple parameters. Dawid and Stone (1982) argue that in order to run the fiducial argument, one has to assume that the statistical problem can be captured in a functional model that is smoothly invertible. In this chapter, I will not discuss the fiducial argument and explore another non-ampliative representation instead.

Note that the probability of the data only depends on the number of 0's and the number of 1's in the sequence. Now the above likelihood function is maximal at $\theta = \frac{1}{3}$, so $\hat{\theta} = \frac{1}{3}$. More generally, defining t_1 as the number of 1's in the sequence s_t, the maximum likelihood estimator is

$$\hat{\theta}(s_t) = \frac{t_1}{t}. \tag{5}$$

Note finally that for a true value θ, the probability of finding the estimate in the interval $\frac{t_1}{t} \in [\theta - \Delta, \theta + \Delta]$ increases for larger data sequences. Fixing the probability at $1 - \epsilon$, the size of the interval will therefore decrease with increasing sample size.

This completes the introduction into parameter estimation. The thing to note is that the statistical procedure can be taken as the procedural analogue of an ampliative logical inference, running from the data to a probability assignment over the sample space. We have $\mathcal{H}$ as the set of probability models from which the inference starts, and by means of the data we then choose a single $h_{\hat{\theta}}$ of these, or a set C_{95}, as our conclusion. In the following I aim to investigate whether there is a non-ampliative logical representation of this inductive inference.

6 Carnapian logics

A straightforward way of capturing parameter estimation in a logic is by relating it to the logic of induction developed by Carnap (1950, 1952). Historically, Carnapian inductive logic can lay most claim to the title of inductive logic proper. It was the first systematic study into probabilistic predictions on the basis of data.

The central concept in Carnapian inductive logic is logical probability. Recall that the sample space $\mathcal{Q}$, also called the observation algebra, corresponds to an observation language, comprising of sentences such as "the second pear is green", or formally, q_2^1. The original idea of Carnap was to derive a probability assignment over the language on the basis of symmetries within the language. In the example, we have three mutually exclusive properties for each pear, and in the absence of any further knowledge, there is no reason to think of any of these properties as special or as more, or less, appropriate than the other two. The symmetry inherent to the language suggests that each of the sentences q_i^k for $k = 0, 1, 2$ should get equal probability:

$$P(q_i^0) = P(q_i^1) = P(q_i^2) = \frac{1}{3}.$$

The idea of logical probability is to fix a unique probability function over the observation language, or otherwise a strongly restricted set of such functions, on the basis of symmetries.

Next to symmetries, the set of probability functions can also be restricted by certain predictive properties. As an example, we may feel that yellow pears are more akin to green pears, so that finding a yellow pear decreases the probability for red pears considerably, while it decreases the probability for green pears much less dramatically. That is,

$$\frac{P(q^1_{t+1}|s_{t-1} \cap q^2_t)}{P(q^0_{t+1}|s_{t-1} \cap q^2_t)} \quad > \quad \frac{P(q^1_{t+1}|s_{t-1})}{P(q^0_{t+1}|s_{t-1})}.$$

How such relations among properties may play a part in determining the probability assignment P is described in the literature on analogy reasoning. See Festa (1996); Maher (2000); Romeijn (2006). Interesting recent work on relations between predictive properties in the context of analogical predictions can also be found in Paris and Waterhouse (2008).

Any Carnapian inductive logic is defined by a number of symmetry principles and predictive properties, determining a probability function, or a set of such functions. One very well-known inductive logic, discussed at length in Carnap (1952), employs a probability assignment characterised by the following symmetries,

$$\begin{aligned} P(q^k_i) &= P(q^{k'}_i), \\ P(s^{k_1 \ldots k_i \ldots k_t}) &= P(s^{k_i \ldots k_1 \ldots k_t}), \end{aligned} \tag{6}$$

for all values of i, t, k, and k', and for all values k_i with $1 \leq i \leq t$. The latter of these is known as the exchangeability of observations: the order in the observations does not matter to their probability. The inductive logic at issue employs a particular version of exchangeability, known as the requirement of restricted relevance,

$$P(q^k_{t+1}|s_t) \quad = \quad f(t_k, t), \tag{7}$$

where t_k is the number of earlier instances q^k_i in the sequence s_t and t the total number of observations. Together these symmetries and predictive properties determine a particular set of probability assignments P, for which we can derive the following consequence:

$$P(q^k_{t+1}|s_t) \quad = \quad \frac{t_k + \frac{\lambda}{n}}{t + \lambda}, \tag{8}$$

where n the number of values for k. The parameter $0 \leq \lambda \leq \infty$ can be chosen at will. Predictive probability assignments of this form are called Carnapian λ-rules.

The probability distribution in Equation (8) has some striking features. Most importantly, for any of the probability functions P satisfying the aforementioned symmetries, we have that

$$P(q^k_{t+1}|s_{t-1} \cap q^k_t) \quad > \quad P(q^k_{t+1}|s_{t-1}).$$

This predictive property is called instantial relevance: the occurrence of q^k_t increases the probability for q^k_{t+1}. It was a success for Carnap that this typically inductive effect is derivable from the symmetries alone. By providing an independent justification for these symmetries, Carnap effectively provided a justification for induction, thereby answering the age-old challenge of Hume.[6]

Note that the outlook of Carnapian logic is very different from the outlook of classical statistical procedures, like Fisher's parameter estimation or Neyman-Pearson testing. Classical statistics starts with statistical hypotheses, each associated with a probability functions over a sample space, and then chooses the best fitting one on the basis of the data. By contrast, Carnapian logic starts with a sample space and a number of symmetry principles and predictive properties, that together fix a set of probability functions over the sample space. Just like the truth tables restrict the possible truth valuations, so do these principles restrict the logical probability functions, albeit not to a singleton, as λ can still be chosen freely. But from the point of view of statistics, Carnap is thereby motivating, from logical principles, the choice for a particular set of hypotheses.

Recall that classical statistics was naturally associated with ampliative inductive inference. By contrast, if we ignore the notion of logical probability and concentrate on the inferential step, Carnapian inductive logics fall very neatly within the template for non-ampliative inductive logic that I laid down at the beginning. By means of a number of symmetry principles and predictive properties, we fix a set of probability assignments over the sample space. The conclusions are then reached by working out specific consequences for probability functions within this set, using the axioms of probability. In particular, Carnapian inductive logic looks at the probability assignments conditional on various samples s_t, deriving that they all satisfy instantial relevance, for example. Importantly, the symmetries in the lan-

[6] As recounted in Zabell (1982), earlier work that connects exchangeability to the predictive properties of probability functions was done by Johnson (1932) and de Finetti (1937). But the specific relation with Hume's problem noted here is due to Carnap: he motivated predictive properties such as Equation (8) independently, by the definition of logical probability, whereas for the subjectivist de Finetti these properties did not have any objective grounding.

guage appear as premises in the inductive logical inference. They restrict the set of probability assignments that is considered in the inference.

Despite these differences in outlook, ampliative against non-ampliative, we can identify a strong similarity between parameter estimation, as discussed in Section 5, and the predictive systems of Carnapian logic. To see this, note that the procedure of parameter estimation can be used to determine the probability of the next piece of data. In the example on pears, once we have observed s^{000101} and thus chosen $h_{\frac{1}{3}}$, we may on the basis of that predict that the next pear has a probability of $\frac{1}{3}$ to be green. In other words, the function $\hat{\theta}$ is a predictive system, much like any other Carnapian inductive logic. We can write

$$P(q^k_{t+1}|s_t) \;\; = \;\; P_{h_{\hat{\theta}(s_t)}}(q^k_{t+1}).$$

The estimation function $\hat{\theta}$ by Fisher is thus captured in a single probability function P. So we can present the latter as a probability assignment over sample space, from which estimations can be derived by a non-ampliative inference.

Let me make this concrete by means of the example on red and green pears. In the Carnapian prediction rule of Equation (8), choosing $\lambda = 0$ will yield the observed relative frequency as predictions. And according to Equation (5) these relative frequencies are also the maximum likelihood estimators. Thus, for each set of possible observations, $\{s^{k_1 \dots k_t} : k_i = 0, 1\}$, the Carnapian rule with $\lambda = 0$ predicts according to the Fisherian estimate.[7]

Unfortunately the alignment of Fisher estimation and Carnapian inductive logic is rather problematic. Already for estimations for multinomial hypotheses, it is not immediate how we can define the corresponding probability assignment over sample space, and whether we thereby define a coherent probability function at all. For more complicated sets of hypotheses, and the more complicated estimators associated with it, the corresponding probability assignment P may be even less natural, or possibly incoherent. Moreover, the principles and predictive properties that may motivate

[7] Note that the probability function P that describes the estimations is a rather unusual one. After three green pears for example, s^{111}, the probability for the next pear to be red will be 0, so that $P(s^{1110}) = 0$. By the standard axiomatisation and definitions of probability, the probability of any observation q^0_5 conditional on s^{1110} is not defined. But if the probability function P is supposed to follow the Fisherian estimations, then we must have $P(q^0_5|s^{1110}) = \frac{1}{4}$. To accommodate the probability function imposed by Fisher's estimations, we may change the axiomatisation of probability. In particular, we may adopt an axiomatisation in which conditional probability is primitive, as described in Rényi (1970). Alternatively, we can restrict ourselves to estimations based on the observation of more than one property.

the choice of that probability function will be very hard to come by. In the following I will therefore not discuss the further intricacies of capturing Fisher's estimation functions by Carnapian prediction rules. However, Carnapian rules will make a reappearance in the next two sections, because in a much more straightforward sense, they are the predictive counterpart to Bayesian statistics.

7 Bayesian statistics

The defining characteristic of Bayesian statistics is that probability assignments do not just range over data, but that they can also take statistical hypotheses as arguments. As will be seen in the following, Bayesian inference is naturally represented in terms of a non-ampliative inductive logic, and it also relates very naturally to Carnapian inductive logic.

Let $\mathcal{H}$ be the space of statistical hypotheses h_θ, and let $\mathcal{Q}$ be the sample space as before. The functions P are probability assignments over the entire space $\mathcal{H} \times \mathcal{Q}$. Since h_θ is a member of the combined algebra, it makes sense to write $P(s_t|h_\theta)$ instead of the $P_{h_\theta}(s_t)$ written in the context of classical statistics. We can define Bayesian statistics as follows.

Definition 7.1 (Bayesian Statistical Inference) *Assume the prior probability $P(h_\theta)$ assigned to hypotheses $h_\theta \in \mathcal{H}$, with $\theta \in \Theta$, the space of parameter values. Further assume $P(s_t|h_\theta)$, the probability assigned to the data s_t conditional on the hypotheses, called the likelihoods. Bayes' theorem determines that*

$$P(h_\theta|s_t) \quad = \quad P(h_\theta)\frac{P(s_t|h_\theta)}{P(s_t)}. \tag{9}$$

Bayesian statistics outputs the posterior probability assignment, $P(h_\theta|s_t)$.

See Barnett (1999) and Press (2003) for a more detailed discussion. The further results form a Bayesian inference, such as estimations and measures for the accuracy of the estimations, can all be derived from the posterior distribution over the statistical hypotheses.

In this definition the probability of the data $P(s_t)$ is not presupposed, because it can be computed from the prior and the likelihoods by the law of total probability,

$$P(s_t) \quad = \quad \int_\Theta P(h_\theta)P(s_t|h_\theta)d\theta.$$

The result of a Bayesian statistical inference is not always a posterior probability. Often the interest is only in comparing the ratio of the posteriors of two hypotheses. By Bayes' theorem we have

$$\frac{P(h_\theta|s_t)}{P(h_{\theta'}|s_t)} \;=\; \frac{P(h_\theta)P(s_t|h_\theta)}{P(h_{\theta'})P(s_t|h_{\theta'})},$$

and if we assume equal priors $P(h_\theta) = P(h_{\theta'})$, we can use the ratio of the likelihoods of the hypotheses, the so-called Bayes factor, to compare the hypotheses.

Let me give an example of a Bayesian procedure. Consider the hypotheses of Equation (3), concerning the fraction of green pears in Emma's orchard. Instead of choosing among them on the basis of the data, assign a so-called Beta-distribution over the range of hypotheses,

$$P(h_\theta) \;\propto\; \theta^{\lambda/2-1}(1-\theta)^{\lambda/2-1} \tag{10}$$

with $\theta \in \Theta = [0,1]$. For $\lambda = 2$, this function is uniform over the domain. Now say that we obtain a certain sequence of pears, s^{000101}. By the likelihood of the hypotheses as given in Equation (4), we can derive

$$P(h_\theta|s^{000101}) \;=\; \theta^{\lambda/2+1}(1-\theta)^{\lambda/2+3}.$$

More generally, the likelihood function for the data s_t with numbers t_k of earlier instances q_i^k is $\theta^{t_1}(1-\theta)^{t_0}$, so that

$$P(h_\theta|s_t) \;\propto\; \theta^{\lambda/2-1+t_1}(1-\theta)^{\lambda/2-1+t_0}. \tag{11}$$

is the posterior distribution over the hypotheses. This posterior is derived by the axioms of probability theory alone, specifically by Bayes' theorem.

As said, capturing this statistical procedure in a non-ampliative inference is relatively straightforward. The premises are the prior over the hypotheses, $P(h_\theta)$ for $\theta \in \Theta$, and the likelihood functions, $P(s_t|h_\theta)$ over the algebras $\mathcal{Q}$, which are determined for each hypothesis h_θ separately. These premises are such that only a single probability assignment over the space $\mathcal{H} \times \mathcal{Q}$ remains. In other words, the premises have a unique probability model. Moreover, all the conclusions are straightforward consequences of this probability assignment. They can be derived from the assignment by applying theorems of probability theory, primarily Bayes' theorem.

Before turning to the relation of Bayesian inference with Carnapian logic, let me compare it to the classical procedures sketched in the foregoing. In all cases, we consider a set of statistical hypotheses, and in all cases our choice among these is informed by the probability of the data according to the hypotheses. The difference is that in the two classical procedures, this choice is absolute: acceptance, rejection, and the appointment of a best estimate. In the Bayesian procedure, by contrast, all this is expressed in a posterior probability assignment over the set of hypotheses.

Note that this posterior over hypotheses can be used to generate the kind of choices between hypotheses that classical statistics provides. Consider Fisherian parameter estimation. We can use the posterior to derive an expectation for the parameter θ, as follows:

$$\mathrm{E}[\theta] \quad = \quad \int_{\Theta} \theta P(h_{\theta}|s_t) d\theta. \tag{12}$$

Clearly, $\mathrm{E}[\theta]$ is a function that brings us from the hypotheses h_θ and the data s_t to a preferred value for the parameter. The function depends on the prior probability over the hypotheses, but it is in a sense analogous to the maximum likelihood estimator. In analogy to the confidence interval, we can also define a so-called credal interval from the posterior probability distribution:

$$\mathrm{Cred}_{1-\epsilon} \quad = \quad \left\{ \theta : \, |\theta - \mathrm{E}[\theta]| < d \ \text{ and } \int_{\mathrm{E}[\theta]-d}^{\mathrm{E}[\theta]+d} P(h_\theta|s_t) d\theta = 1 - \epsilon \right\}.$$

This set of values for θ is such that the posterior probability of the corresponding h_θ jointly add up to $1 - \epsilon$ of the total posterior probability.

Most of the controversy over the Bayesian method concerns the determination and interpretation of the probability assignment over hypotheses. As for interpretation, classical statistics objects to the whole idea of assigning probabilities to hypotheses. The data have a well-defined probability, because they consist of repeatable events, and so we can interpret the probabilities as frequencies, or as some other kind of objective probability. But the probability assigned to a hypothesis cannot be understood in this way, and instead expresses an epistemic state of uncertainty. One of the distinctive features of classical statistics is that it rejects such epistemic probability assignments, and that it restricts itself to a straightforward interpretation of probability as relative frequency.

Even if we buy into this interpretation of probability as epistemic uncertainty, how do we determine a prior probability? At the outset we do not have any idea of which hypothesis is right, or even which hypothesis is a good candidate. So how are we supposed to assign a prior probability to the hypotheses? The literature proposes several objective criteria for filling in the priors, for instance by maximum entropy or by other versions of the principle of indifference, but something of the subjectivity of the starting point remains. The strength of the classical statistical procedures is that they do not need any such subjective prior probability.

8 Bayesian inductive logic

While Bayesian statistics differs strongly from classical statistics, it is much more closely related to the inductive logic of Carnap. In this section I will elaborate on this relation, and indicate how Bayesian statistical inference and inductive logic may have a fruitful common future.

To see how Bayesian statistics and Carnapian inductive logic hang together, note first that the result of a Bayesian statistical inference, namely a posterior, is naturally translated into the result of a Carnapian inductive logic, namely a prediction,

$$P(q^1_{t+1}|s_t) = \int_0^1 P(q^1_{t+1}|h_\theta \cap s_t)P(h_\theta|s_t)d\theta, \tag{13}$$

by the law of total probability. Furthermore, consider the posterior probability over multinomial hypotheses. Recall that the parameter θ is the probability for the next pear to be green, as defined in Equation (3). By Equations (12) and (13) we have

$$\begin{aligned} \mathrm{E}[\theta] &= \int_\Theta \theta P(h_\theta|s_t)d\theta \\ &= \int_0^1 P(q^1_{t+1}|h_\theta \cap s_t)P(h_\theta|s_t)d\theta \\ &= P(q^1_{t+1}|s_t), \end{aligned}$$

This shows that in the case of multinomial statistical hypotheses, the expectation value for the parameter is the same as a predictive probability.

The correspondence between Bayesian statistics and Carnapian inductive logic is in fact even more striking. We can work out the integral of Equation line (13), using Equation (10) as the prior and the multinomial hypotheses defined in Equation 3, to obtain

$$P(q^1_{t+1}|s_t) \;=\; \frac{t_1 + \frac{\lambda}{2}}{t + \lambda}. \tag{14}$$

This means that there is a specific correspondence between certain kinds of predictive probabilities, as described by the Carnapian λ-rules, and certain kinds of Bayesian statistical inferences, namely with multinomial hypotheses and priors from the family of Dirichlet distributions, which generalise the Beta-distributions used in the foregoing.

On top of this, the equivalence between Carnapian inductive logic and Bayesian statistical inference is more general than is shown in the foregoing. Instead of the well-behaved priors just considered, we might consider any functional form as a prior over the hypotheses h_θ, and then wonder what the resulting predictive probability is. As de Finetti showed in his representation theorem, the resulting predictive probability will always comply to a predictive property known as exchangeability, which was given in Equation (6). Conversely, and more surprisingly, any predictive probability complying to the property of exchangeability can be written down in terms of a Bayesian statistical inference with multinomial hypotheses and some prior over these hypotheses. In sum, de Finetti showed that there is a one-to-one correspondence between the predictive property of exchangeability on the one hand, and Bayesian statistical inferences using multinomial hypotheses on the other.

It is insightful to make this result by de Finetti explicit in terms of the non-ampliative inductive logic discussed in the foregoing. Recall that a Bayesian statistical inference takes a prior and likelihoods as premises, leading to a single probability assignment over the space $\mathcal{H} \times \mathcal{Q}$ as the only assignment satisfying the premises. We infer probabilistic consequences, such as the posterior and the predictions, from this probability assignment. Similarly, a Carnapian inductive logic is characterised by a single probability assignment, defined over the space $\mathcal{Q}$, from which the predictions can be derived. The representation theorem by de Finetti effectively shows an equivalence between these two probability assignments: when it comes to predictions, we can reduce the probability assignment over $\mathcal{H} \times \mathcal{Q}$ to an assignment over $\mathcal{Q}$ only.

For de Finetti, this equivalence was very welcome. He had a strictly subjectivist interpretation of probability, believing that probability expresses uncertain belief only. Moreover, he was eager to rid science of its metaphysical excess baggage to which, in his view, the notion of objective chance belonged. So de Finetti applied his representation theorem to argue against the use of multinomial hypotheses, and thereby against the use of statistical hypotheses more generally. Why refer to these obscure chances if we can achieve the very same statistical ends by employing the unproblematic notion of exchangeability? The latter is a predictive property, and it can hence be interpreted as an empirical and as a subjective notion.

The fact is that statistics, as it is used in the sciences, is persistent in its use of statistical hypotheses. Therefore I want to invite the reader to consider the inverse application of de Finetti's theorem. Why does science use these obscure objective chances? As I argue extensively in Romeijn (2005), the reason is that statistical hypotheses provide invaluable help by, indirectly, pinning down the probability assignments over $\mathcal{Q}$ that have the required predictive properties. Rather than reducing the Bayesian inferences over statistical hypotheses to inductive predictions over observations, we can use the representation theorem to capture relations between observations in an insightful way, namely by citing the statistical hypotheses that may be true of the data. As further illustrated in Romeijn (2004, 2006), enriching inductive logic in this way improves the control that we have over predictive properties.

Finally, it may be noted that this view on inductive logic is comparable to the "presupposition view" in Festa (1993), which takes a similar line with regards to the choice of λ in Carnapian inductive logic. It is also strongly related to the views expressed by Hintikka in Auxier and Hahn (2006), and I want to highlight certain aspects of this latter view in particular. In response to Kuipers' overview of inductive logic, Hintikka writes that "Inductive inference, including rules of probabilistic induction, depends on tacit assumptions concerning the nature of the world. Once these assumptions are spelled out, inductive inference becomes in principle a species of deductive inference." Now the symmetry principles and predictive properties used in Carnapian inductive logic are exactly the tacit assumptions Hintikka speaks about. As explained in the foregoing, the use of particular statistical hypotheses in a Bayesian inference comes down to the very same set of assumptions, but now these assumptions are not tacit anymore: they have been made explicit as the choice for a particular set of statistical hypotheses. Therefore, the use of statistical hypotheses that I have advertised above may help us to get closer to the ideal of inductive logic envisaged by Hintikka.

9 Neyman-Pearson test as an inference

In this final section, I investigate whether we can turn the Neyman-Pearson procedure of Section 4 into an inference within Bayesian inductive logic. This might come across as a pointless exercise in statistical yoga, trying to make Neyman and Pearson relax in a position that is far from natural. However, the exercise will nicely illustrate the use of Bayesian inductive logic. Moreover, I think that it will bring Neyman-Pearson testing and inductive logic closer together, and thereby stimulate research on the intersection of inductive logic and statistics in the sciences.

An additional reason for investigating Neyman-Pearson hypothesis testing in this framework is that in many practical applications, scientists are tempted to read the probability statements about the hypotheses inversely: the significance is often taken as the probability that the null hypothesis is true. Although emphatically wrong, this inferential reading has a strong intuitive appeal to users. The following will make explicit that in this reading, the Neyman-Pearson procedure is effectively taken as a kind of non-ampliative inductive inference.

First, we construct the space $\mathcal{H} \times \mathcal{Q}$, and define the probability functions P_{h_j} over the sample spaces $\langle h_j, \mathcal{Q} \rangle$. For the prior probability assignment over the two hypotheses, we take $P(h_0) \in (l, u)$, meaning that $l < P(h_0) < u$. Finally, we adopt the restriction that $P(h_0) + P(h_1) = 1$. This defines a set of probability functions over the entire space, serving as a starting point of the inference.

Next we include the data in the probability assignments. Crucially, we coarse-grain the observations to the simple observation f^j, with

$$f^j \quad = \quad \{s_t : F(s_t) = j\},$$

so that the observation simply encodes the value of the test function. It follows from this coarse-graining that we obtain the type-I and type-II errors as the likelihoods of the observations,

$$\begin{aligned} P(f^1|h_0) &= \alpha, \\ P(f^0|h_1) &= \beta. \end{aligned}$$

Finally we use Bayes' theorem to derive a set of posterior probability distributions over the hypotheses, according to

$$\frac{P(h_1|f^j)}{P(h_0|f^j)} \quad = \quad \frac{P(f^j|h_1)P(h_1)}{P(f^j|h_0)P(h_0)}.$$

Note that the quality of the test, in terms of size and power, will be reflected in the posteriors. If, for example, we find an observation s_t that allows us to reject the null hypothesis, so f^1, then as long as $\alpha < 1 - \beta$, meaning that the significance is smaller than the power, we find that $\underline{P}(h_0|f^1) < \underline{P}(h_0)$ and $\overline{P}(h_0|f^1) < \overline{P}(h_0)$. The larger the difference between significance and power, the larger the difference between posteriors and priors.

Note, however, that we have not yet decided on a fully specified prior probability over the statistical hypotheses. This echoes the fact that classical statistics does not make use of a prior probability. However, it is only by restricting the prior probability over hypotheses in some way or other that we can make the Bayesian rendering of the results of Neyman and Pearson work. In particular, if we choose $(l, u) = (0, 1)$ for the prior, then we find $(l', u') = (0, 1)$ for the posterior as well. However, if we choose

$$l \;\geq\; \frac{\beta}{\beta + 1 - \alpha}, \qquad u \;\leq\; \frac{1 - \beta}{1 - \beta + \alpha},$$

we find for all $P(h^0) \in (l, u)$ that $\overline{P}(h_0|f^1) < \frac{1}{2} < \underline{P}(h_1|f^1)$. Similarly, we find $\underline{P}(h_0|f^0) > \frac{1}{2} > \overline{P}(h_1|f^0)$. So with this interval prior, an observation s_t for which $F(s_t) = 1$ tilts the balance towards h_1 for all the probability functions P in the interval, and vice versa.

Let me illustrate the Bayesian inference by means of the above example on pears. We set up the sample space and hypotheses as before, and we then coarse-grain the observations to f^j, corresponding to the value of the test function, $f^1 = q^0 \cup q^1$ and $f^0 = q^2$. We obtain

$$\begin{array}{rclclcl} P(f^1|h_0) & = & P(q^0 \cup q^1|h_0) & = & \alpha & = & 0.05 \\ P(f^0|h_1) & = & P(q^0 \cup q^1|h_1) & = & \beta & = & 0.30 \end{array}$$

Choosing $P(h_0) \in (0.24, 0.93)$, this results in $P(h_0|f^0) = (0.50, 0.98)$, and $P(h_0|f^1) = (0.02, 0.50)$.

Depending on the choice of prior, one can argue that the resulting Bayesian inference replicates the Neyman-Pearson procedure: if the probability over hypotheses expresses our preference over them, then indeed f^0 makes us prefer h_0 and f^1 makes us prefer h_1. Importantly, the inference fits the entailment relation mentioned earlier: we have a set of probabilistic models on the side of the premises, namely the set of priors over $\mathcal{H}$, coupled to the full probability assignments over $\langle h_j, \mathcal{Q} \rangle$ for $j = 0, 1$. And we have a set of models on the conclusion side, namely the set of posteriors over $\mathcal{H}$. Because the latter is computed from the former by the axioms of probability, the two sets include the same probability functions. Therefore the conclusion is classically entailed by the premises.

The above example shows that we can imitate the workings of a Neyman-Pearson test in Bayesian inductive logic, and thus in terms of a non-ampliative inductive inference. But the imitation is far from perfect. For one, the results of a Bayesian inference will always be a probability function. By contrast, Neyman-Pearson statistics ends in a decision to accept or reject, which is a binary decision instead of some sort of weak or inconclusive preference. Of course, there are many attempts to weld a binary decision onto the probabilistic end result of a Bayesian inference, for example in Levi (1980) and in the discussion on rational acceptance, e.g., Douven (2002). In particular, we might supplement the probabilistic results of a Bayesian inference with rules for translating the probability assignments into decisions, e.g., we choose h_0 if we have $\underline{P}(h_0|s_t) > \frac{1}{2}$, and similarly for h_1. However, the bivalence of Neyman-Pearson statistics cannot be replicated in a Bayesian inference itself. It will have to result from a decision-theoretic add-on to the inferential part of Bayesian statistics.

More generally, the representation in probabilistic logic will probably not appeal to advocates of classical statistics. Quite apart from the issue of binary acceptance, the whole idea of assuming a prior probability, however unspecific, may be objected to on the principled ground that probability functions express long-term frequencies, and that hypotheses cannot have such frequencies.

There is one attractive feature, at least to my mind, of the above rendering, that may be of interest in its own right. With the representation in place, we can ask again how to understand the example by Jeffreys, as considered in Section 4. Following Edwards (1972), it illustrates that Neyman and Pearson tests do not respect the likelihood principle, because they depend on the probability assignment over the entire sample space and not just on the probability of the observed sample. However, in the Bayesian representation we do respect the likelihood principle, but in addition we condition on f^j, not on q^k. In fact the whole example hinges on how the samples are grouped into regions of acceptance and rejection. Instead of adopting the diagnosis by Hacking concerning the likelihood principle, we could therefore say that the approach of Neyman and Pearson takes the observations in terms of a rather coarse-grained partition of information. In other words, rather than saying that Neyman-Pearson procedures violate the likelihood principle, we can also say that the procedures crucially depend on how the observed sample is framed, and thus violate the principle of total evidence.

10 In conclusion

In the foregoing I have discussed three statistical procedures, to wit, Neyman-Pearson hypotheses testing, Fisher's maximum likelihood estimation, and Bayesian statistical inference. These three procedures were seen to relate to inductive logic in a variety of ways.

The two classical approaches were connected most naturally to ampliative inductive inference, running from a set of probability functions and the data to a restricted set of such functions. However, I have also related both procedures to non-ampliative inferences. First I connected parameter estimation to Carnapian inductive logic. Then I related this logic to Bayesian statistical inference, which was seen to be non-ampliative already. Further, I have indicated how Carnapian inductive logic can be extended to Bayesian inductive logic, which accommodates the use of statistical hypotheses and thus captures Bayesian statistics. Finally, I have illustrated the latter logic by giving a non-ampliative account of Neyman-Pearson hypothesis testing.

I hope that portraying statistical procedures in the setting of inductive logic has been illuminating. In particular, I hope that the relation between Carnapian inductive logic and Bayesian statistics stimulates research on the intersection of the two. Certainly, some research in this area has already been conducted; see for example Skyrms (1991, 1993, 1996) and Festa (1993). Following these contributions, Romeijn (2005) argues that an inductive logic that includes statistical hypotheses in its language is closely related to Bayesian statistical inference, and some of these views have been reiterated in this chapter. However, I believe that there is much room for improvement. Research on the intersection of inductive logic and statistical inference can certainly enhance the relevance of inductive logical systems to scientific method and the philosophy of science. In parallel, I believe that insights from inductive logic may help to clarify the foundations of statistics.

Acknowledgements

I want to thank Roberto Festa for comments on an earlier draft of this chapter, and Jos Uffink for drawing my attention to Jeffreys' amusing quote on the "remarkable procedure". I also thank the Spanish Ministry of Science and Innovation (Research project FFI2008-1169) for financial support. This research was carried out as part of a project funded by the Dutch Organization of Scientific Research (NWO VENI-grant nr. 275-20-013).

References

Auxier, R. and Hahn, L., editors (2006). *The Philosophy of Jaako Hintikka.* Open Court, Chicago.

Barnett, V. (1999). *Comparative Statistical Inference.* John Wiley, New York.

Carnap, R. (1950). *Logical Foundations of Probability.* University of Chicago Press.

Carnap, R. (1952). *The Continuum of Inductive Methods.* University of Chicago Press, Chicago.

Dawid, A. P. and Stone, M. (1982). The functional-model basis of fiducial inference (with discussion). *Annals of Statistics*, 10(4):1054–1074.

de Finetti, B. (1937). La prévision: ses lois logiques, ses sources subjectives. *Annales de l'Institut Henri Poincaré*, 7(1):1–68.

Douven, I. (2002). A new solution to the paradoxes of rational acceptability. *The British Journal for the Philosophy of Science*, 53:391–410.

Edwards, A. (1972). *Likelihood.* Cambridge University Press.

Festa, R. (1993). *Optimum Inductive Methods.* Dordrecht: Kluwer.

Festa, R. (1996). Analogy and exchangeability in predictive inferences. *Erkenntnis*, 45:89–112.

Fisher, R. A. (1930). Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26:528–535.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*, 6:317–324.

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

Hacking, I. (1965). *The Logic of Statistical Inference.* Cambridge University Press, Cambridge.

Haenni, R., Romeijn, J.-W., Wheeler, G., and Williamson, J. (2009). *Probabilistic Logics and Probabilistic Networks.* Springer.

Hailperin, T. (1996). *Sentential Probability Logic.* Lehigh University Press.

Hartmann, S., Gabbay, D., and Woods, J., editors (2009). *Handbook for the History of Logic: Inductive Logic (Volume 10).* College Publications.

Howson, C. (2003). Probability and logic. *Journal of Applied Logic*, 1(3–4):151–165.

Jeffreys, H. (1931). *Scientific Inference.* Cambridge University Press, , Cambridge.

Johnson, W. (1932). Probability: the deductive and inductive problems. *Mind*, 49:409–423.

Kyburg, Jr., H. E. (1974). *The Logical Foundations of Statistical Inference.* D. Reidel, Dordrecht.

Levi, I. (1980). *The enterprise of knowledge: an essay on knowledge, credal probability, and chance.* MIT Press, Cambridge MA.

Maher, P. (2000). Probabilities for two properties. *Erkenntnis*, 52:63–81.

Neyman, J. and Pearson, E. (1967). *Joint Statistical Papers.* University of California Press, Berkeley.

Paris, J. and Waterhouse, P. (2008). Atom exchangeability and instantial relevance. *unpublished manuscript.*

Press, J. (2003). *Subjective and Objective Bayesian Statistics: Principles, Models, and Applications.* John Wiley, New York.

Rényi, A. (1970). *Probability Theory.* North Holland, Amsterdam.

Romeijn, J. (2004). Hypotheses and inductive predictions. *Synthese*, 141(3):333–64.

Romeijn, J. (2005). *Bayesian Inductive Logic.* PhD dissertation, University of Groningen.

Romeijn, J. (2006). Analogical predictions for explicit similarity. *Erkenntnis*, 64:253–280.

Seidenfeld, T. (1979). *Philosophical Problems of Statistical Inference: Learning from R.A. Fisher.* Reidel, Dordrecht.

Skyrms, B. (1991). Carnapian inductive logic for markov chains. *Erkenntnis*, 35:35–53.

Skyrms, B. (1993). Analogy by similarity in hyper-Carnapian inductive logic. In Earman, J., Janis, A. I., Massey, G., and Rescher, N., editors, *Philosophical Problems of the Internal and External Worlds*, pages 273–282. University of Pittsburgh Press, Pittsburgh.

Skyrms, B. (1996). *Statistics, Probability, and Game*, chapter Carnapian Inductive Logic and Bayesian Statistics, pages 321–336. IMS Lecture Notes.

Wheeler, G. (2006). Rational acceptance and conjunctive/disjunctive absorption. *Journal of Logic, Language and Information*, 15(1-2):49–63.

Zabell, S. (1982). W. e. johnson's “sufficientness” postulate. *Annals of Statistics*, 10:1091–99.